The Employee Promotion Decision based on the Randomforest Algorithm and the Analytic Hierarchy Process

Yanming Chen[1,*], Xinyu Lin[2a], Kunye Zhan[2b]

[1]Student, Shantou University, Shantou, China
[2]Student, South China Normal University, Guangzhou, China
[2]Student, Shenzhen University, Shenzhen, China

*21ymchen@stu.edu.cn,
[a]xinyul2002@163.com, [b]2021040486@email.szu.edu.cn

Abstract. This paper aims to build an employee promotion decision model based on the Randomforest algorithm and the Analytic Hierarchy Process. Random Undersampling algorithm is applied to resolve the issue of data imbalance and the Point-biserial analysis is employed for the paper to conduct feature filtering after data cleaning and preprocessing. Subsequently, we employ the Randomforest algorithm to establish a classification model for employee promotion, alongside a logistic regression algorithm for comparative purposes. Ultimately, we optimize the decision-making system using the Analytic Hierarchy Process (AHP) to improve its overall efficiency.This model holds significant implications for both employee promotion decision and human resource management.

Keywords: Employee Promotion; Randomforest; Analytic Hierarchy Process; Point-biserial; Human Resource Management

1 Introduction

Employee promotion decision is an important research topic in human resource management, and lots of scholars have attempted to establish employee promotion prediction models, such as "Imbalanced Employee Promotion Prediction Based on XGBoost" [1], "Exploring Incentive Promotion Mechanisms for Enterprise Employees" [2], "Optimization Strategies for Enterprise Employee Promotion Management Systems" [3], and "A Brief Discussion on the Application of Employee Points-based Promotion Incentive Models in Enterprises" [4]. However, the models built by widely used machine learning algorithms often struggle to be effectively applied in practical decision-making scenarios.

This paper proposes a model that combines the Randomforest algorithm with the analytic hierarchy process (AHP) [5], providing a comprehensive reference for employee promotion decisions. It leverages the advantages of machine learning in handling complex relationships and the strengths of AHP in determining the relative importance of

S. Yacob et al. (eds.), *Proceedings of the 2023 7th International Seminar on Education, Management and Social Sciences (ISEMSS 2023)*, Advances in Social Science, Education and Humanities Research 779,
https://doi.org/10.2991/978-2-38476-126-5_185

different criteria. This model can assist decision-makers in making accurate and fair promotion decisions, thereby improving the efficiency and effectiveness of the human resource management process.

2 Theoretical Foundation

Machine learning is a subfield of artificial intelligence that focuses on the development and study of algorithms and models that enable computers to learn from data and make predictions and decisions [6]. Random forest classification algorithm is a typical machine learning algorithm. It combines multiple classification trees to improve prediction accuracy [7]. It uses bagging method to generate multiple classification trees with different training data subsets, and then aggregates the predictions of these trees.

Analytic Hierarchy Process (AHP) is a typical evaluation model in mathematical modeling. It decomposes a problem into different constituent factors based on the nature of the problem and the overall objective to be achieved. These factors are then aggregated and combined into a multi-level analytical hierarchical structure model, taking into consideration their interrelationships and hierarchical dependencies.

3 Materials and Methods

In this paper, variables are first divided into numerical variables and categorical variables. After data cleaning and preprocessing, numerical variables are scaled and categorical variables are transformed into dummy variables. Then, Point-biserial analysis method is employed for feature filtering to reduce the risk of overfitting and the Random Undersampling algorithm is applied to address data imbalance. Afterward, a Randomforest algorithm is used to establish an employee promotion prediction model, which is then compared with a logistic regression model. Finally, the decision system is optimized and improved using the Analytic Hierarchy Process.

3.1 Dataset used in the study

In order to establish a model for employee promotion decision-making, we need to analyze historical dataset. Therefore, this paper selects a large multinational company employee promotion historical dataset from kaggle.com for research purposes. Kaggle.com is a data science platform that provides various datasets for machine learning, data analysis, and data visualization. This dataset consists of 10,819 observations, where each observation represents whether the employee was promoted or not, and also includes other information about the employee. The dataset consists of both numerical and categorical variables, and the ratio between numerical and categorical variables is shown in **Table 1.**

Table 1. The proportion of two categories of variables

Type	Number	Proportion
Categorical	4	36.36%
Numerical	11	63.64%

There are a total of 11 variables in this dataset, with "is promoted" as the dependent variable, and the other 10 columns as independent variables. The basic information of the numerical and categorical variables in the independent variables is shown in **Table 2** and **Table 3**, respectively.

Table 2. Basic information about numerical variables

	Count	Min	Max	Mean
Age	10819	20.00	60.00	34.60
No_of_trainings	10819	1.00	9.00	1.24
Previous_year_rating	10041	1.00	5.00	3.58
Length_of_service	10819	1.00	34.00	5.83
Awards	10819	0.00	1.00	0.06
Avg_training_score	10819	41.00	99.00	66.34

Table 3. Rudimentary information about categorical variables

	Count	Unique	Top	Freq
Department	10819	9	Sales & Marketing	3127
Education	10418	3	Bachelor	7145
Gender	10819	2	male	7500
Recruiment channel	10819	3	other	5956

The variable "No_of_trainings" represents the employee's training time, "Previous_year_rating" represents the employee's performance last year, "Length_of_service" represents the employee's total working time, "Awards" represents whether the employee has received awards before, and "Avg_training_score" represents the employee's average score during training. The distribution of some variables are shown in **Figure 1.**

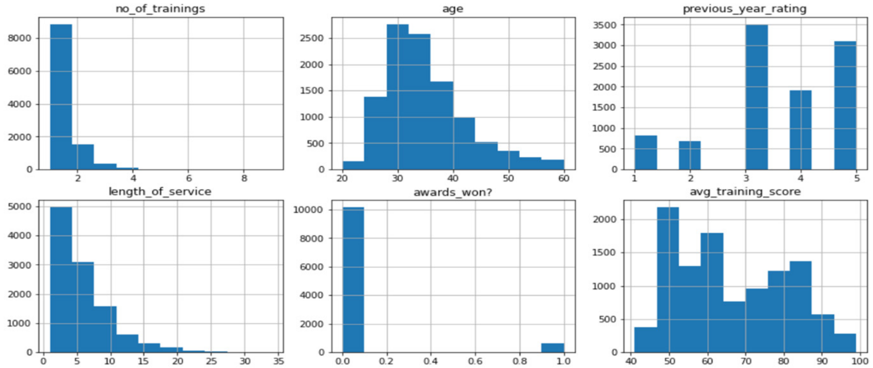


Fig. 1. The distribution of some numerical variables

3.2 Data cleaning and preprocessing

In this dataset, there are 401 missing values in the "Education" variable and 778 missing values in the "Previous_year_rating" variable. Due to the large size of this dataset, we directly delete these missing values, resulting in 10,041 remaining observations. For the outliers in the dataset, which are values that exceed three times the standard deviation or are clearly erroneous, we fill the missing values with the mean value for numerical variables and with the mode value for categorical variables. Then, we use the min-max scaling method to scale the numerical variables, scaling them to the range of 0 to 1, in order to reduce data errors. The formula for "min-max rescaling" is as follows (1):

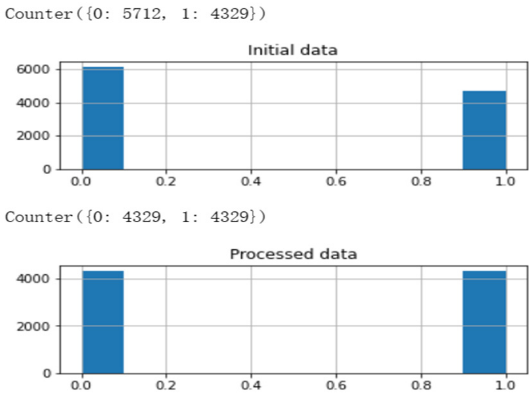
$$X' = \frac{X-Xmin}{Xmax-Xmin} \qquad (1)$$

Where X is the initial value, $Xmax$ and $Xmin$ are the maximum and minimum values, and X' is the final transformed value.

Finally, all categorical variables are converted into dummy variables which are 0-1 variables using one-hot encoding.

3.3 Undersampling algorithm

In this dataset, the dependent variable "is promoted" is represented by "1" for those who have promoted and "0" for those who have not. However, the proportion of "1" and "0" is imbalanced, which may lead to errors if a classification model is directly established. After data cleaning and preprocessing, the number of observations with values 1 and 0 is 4329 and 5712, respectively. To address the issue of data imbalance, we have two methods: oversampling and undersampling algorithms [8]. In this dataset, the oversampling algorithm is to increase the count of 1 to 5712, while the undersampling algorithm is to decrease the count of 0 to 4329. Considering the large size of this dataset and to reduce the risk of overfitting, we use the random undersampling approach from the undersampling algorithm to handle the data. The data before and after undersampling is shown in **Figure 2.**



```
Counter({0: 5712, 1: 4329})
```

```
Counter({0: 4329, 1: 4329})
```

Fig. 2. The data before and after undersampling

3.4 Point-biserial analysis for feature filtering

In this paper, Point-biserial correlation analysis is performed to examine the relationship between all variables and the target variable "attrition". Point-biserial correlation analysis measures the correlation between variables, and the p-value indicates the level of significance. Variables that have a p-value less than 0.05 are generally considered to have no significant correlation with the target variable. Therefore, we have eliminated the variables with a p-value less than 0.05, which are illustrated in **Table 4.**

Table 4. The variables with a p-value less than 0.05 in Point-biserial analysis

Variables	Correlation	Pvalue
Below Secondary	-0.016	0.093
Male	-0.014	0.141
Operations	0.011	0.233
Finance	-0.005	0.554
R&D	-0.001	0.872
Sourcing	0.001	0.959

3.5 Model building

Firstly, we employ the typical classification prediction model called logistic regression model as the control experiment. Logistic regression employs gradient descent to minimize the loss function, using the output values of the regression sigmoid function as probabilities for classification. If the probability value exceeds the threshold, it is classified as positive; otherwise, it is classified as negative.

Secondly, we divide the dataset into a training set and a test set with a ratio of 7:3. Subsequently, we develop a Randomforest classification model and conduct preliminary parameter tuning using random grid search. We further fine-tuned the parameters using grid search [9], while also validating the parameters using Bayesian search. Finally, we conclude the accuracy of both classification models on the training and test sets, as well as the F1-score on the test set. Based on these indicators, we conclude that the Randomforest classification model demonstrates better predictive performance.

3.6 Randomforest importance ranking for feature analysis

After establishing the Randomforest classification model, a feature importance ranking is generated, where the sum of the feature importance for all variables is equal to 1 [10]. The partial feature importance ranking generated by the Randomforest classification model is shown in **Table 5**.

Table 5. The top8 features of the Randomforest feature importance ranking

Feature	Importance
Avg_training_score	0.313
Age	0.185
Length_of_ service	0.144
Previous_year_rating	0.118
Awards	0.038
No_of_trainings	0.028
Sale & Marketing	0.027
Master	0.017

3.7 Analytic Hierarchy Process

In practical applications, relying solely on the random forest classification model built using historical data for employee promotion decisions is insufficient. Firstly, historical data can not fully reflect the current situation at the time of decision-making, as both company and market developments change rapidly. Secondly, the output of the Randomforest classification model is binary, indicating whether an employee is promotable or not. It may happen that multiple employees have a promotable output, but there is only one promotion opportunity. In such cases, the model becomes less effective. Therefore, we use the Analytic Hierarchy Process (AHP) to optimize the employee promotion decision-making system, making it more practical and actionable.

Suppose a manager is considering which one of 10 employees should be given a promotion. The Randomforest classification model indicates that Employee A, Employee B, and Employee C are promotable, while the other seven employees are not. Therefore, the problem is transformed into a decision among Employee A, Employee B, and Employee C. In the AHP, we need to determine the goal level, the criteria level, and the alternative level. The goal level is to select one employee for promotion, the alternative level consists of Employee A, Employee B, and Employee C, and the criteria level represents the influencing variables. Based on the feature importance ranking from the Randomforest model, we can identify relevant decision variables such as employee performance (education, performance, and awards), suitability of employee age, employee tenure, and employee potential.

Next, we need to construct a discriminant matrix. The discriminant matrix has the same number of rows and columns, denoted as n. The scale ranges from 1 to 9, where 1 represents equally important, 3 represents slightly important, 5 represents moderately important, 7 represents strongly important, and 9 represents extremely important. 2, 4, 6, 8 lie between the adjacent midpoints of the values mentioned above, and their inverses can indicate the opposite meaning. The numbers in the discriminant matrix are based on the comparisons made by the decision-makers. In this problem, we need to construct five discriminant matrices: the employee performance discriminant matrix, the suitability of employee age discriminant matrix, the employee tenure discriminant matrix, the employee potential discriminant matrix, and the weight discriminant matrix for the four variables. We have constructed the discriminant matrices for employee A, employee B, and employee C based on their information. The employee performance discriminant matrix and the weight discriminant matrix are shown in **Table 6** and **Table 7**, respectively.

Table 6. The employee performance discriminant matrix

	EmployeeA	EmployeeB	EmployeeC
EmployeeA	1	2	3
EmployeeB	1/2	1	2
EmployeeC	1/3	1/2	1

Table 7. The weight discriminant matrix

	Employee performance	Suitability of employee age	Employee tenure	Employee potential
Employee performance	1	3	4	8
Suitability of employee age	1/3	1	2	3
Employee tenure	1/4	1/2	1	2
Employee potential	1/8	1/3	1/2	1

Then, we perform a consistency test with the following steps: calculate the consistency index (CI) using formula (2):

$$CI = \frac{\lambda max - n}{n-1} \tag{2}$$

where λmax represents the maximum eigenvalue of the discriminant matrix.

The corresponding consistency index (RI) can be obtained from a reference table, and for n = 3, RI = 0.52. Finally, calculate the consistency ratio (CR), which is obtained by dividing CI by RI. If CR < 0.1, it indicates that the consistency of the discriminant matrix is acceptable. When the consistency of the discriminant matrix is acceptable, we can find the maximum eigenvector corresponding to the maximum eigenvalue. Then, normalize the eigenvector to obtain the final score results. Decision-making can be based on the final scores of the three employees.

4 Experiments & Results

4.1 Experiment environment

The dataset comes from a public database named kaggle.com. The first experiment was done in python 3.7.0, the second experiment was done in Matlab R2018a, and the configuration of the computer is shown in Table 8.

Table 8. The configuration of the computer

Hardware	Hardware model
CPU	Intel core i7 CPU 2.90 GHZ
RAM	40.0 GB

4.2 Experiments & Results

Firstly, comparative experiment is conducted using logistic regression model. Then, we conducted experiment using the Randomforest classification model, and the results are shown in **Table 9.**

Table 9. Experimental results of classification models

	Training set (accuracy)	Testing set (accuracy)	Testing set(F1-score)
Logistic	1.0	0.88	0.85
Randomforest	1.0	0.93	0.94

The ROC curve of the Randomforest classification model is shown in **Figure 3.**

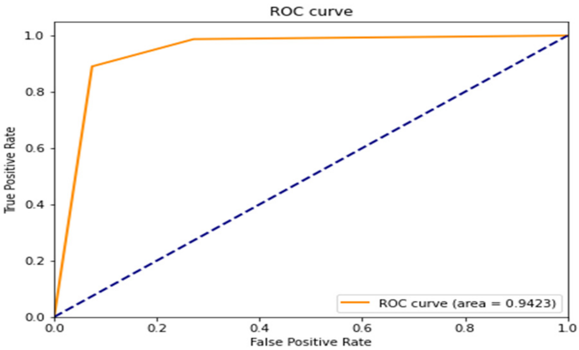


Fig. 3. The ROC curve of the Randomforest classification model

Secondly, we conducted an experiment using the Analytic Hierarchy Process (AHP) on Employee A, B, and C. After passing the consistency test, we obtained the final scores as shown in **Table 10.** From the experimental results, we finally choose Employee A.

Table 10. The final score based on the AHP

Target	Target Weight	A	B	C
Employee performance	0.5786	0.5396	0.297	0.1634
Suitability of employee age	0.2204	0.2857	0.5714	0.1429
Employee tenure	0.1309	0.25	0.5	0.25
Employee potential	0.07	0.2098	0.2402	0.5499
Final score	1	0.422592	0.380045	0.197256

5 Conclusions

In this paper, we propose a model that combines machine learning with the Analytic Hierarchy Process (AHP) for employee promotion decisions. This model can effectively enhance the objectivity and rationality of decision-making. However, considering the complexity and multi-dimensionality of decision problems, the current method lacks sufficient consideration of judgment dimensions and further improvements are needed.

References

1. J. Huang, and H.H. Zhen. (2023) “Imbalanced Employee Promotion Prediction Based on XGBoost,” Software Engineering, 26:25-29. 10.19644/j.cnki.issn2096-1472.2023.003.006.
2. W.Y. Gao. (2020) “Exploring Incentive Promotion Mechanisms for Enterprise Employees,” Human Resources, 22:144-145. https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C44YLTlOAiTRKibYlV5Vjs7iy_Rpms2pqwbFRRUtoUImHcehB-ZXKZ_sf4i_cOF197r66x0Di_fys9jjvh4SdYsh&uniplatform=NZKPT&src=copy.
3. J.B. Yang. (2022) “Optimization Strategies for Enterprise Employee Promotion Management Systems,” Modern Enterprise Culture, 27: 142-144. https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C44YLlOAiTRKibYlV5Vjs7ioT0BO4yQ4m_mOgeS2ml3UBhVqLrEihdD3c6g0ZN_V4KCO19-n3ZnuGOHrbBdqBak&uniplatform=NZKPT&src=copy.
4. H.J. Ju. (2023) “A Brief Discussion on the Application of Employee Points-based Promotion Incentive Models in Enterprises,” Market Modernization, 05: 102-104. 10.14013/j.cnki.scxdh.2023.05.047.
5. C.P. Xie. (2018) “The application of the Analytic Hierarchy Process in employee promotion and selection,” Times Finance, 33: 297-298. https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C44YLTlOAiTRKibYlV5Vjs7iLik5jEcCI09uHa3oBxtWoGV33w1ECd0e4lVUpQf3NAk-BQwnyxwucB5GJBAD4D_bD&uniplatform=NZKPT&src=copy.

6. N.A. Jalil, H.J. Hwang , and N.M. Dawi. (2019) “ Machines Learning Trends, Perspectives and Prospects in Education Sector,” In: Education and Multimedia Technology 2019. pp.201-205. 10.1145/3345120.3345147.
7. X.Y. Liu. (2023) “A study on the applicability of the random forest algorithm in monitoring watershed area,” Modern Information Technology, 7: 74-77. 10.19850/j.cnki.2096-4706.2023.12.019.
8. H. Xiao, L.L. Li. (2022) “Research on credit risk based on the random undersampling algorithm,” Journal of Qingdao University(Natural Science Edition), 35: 126-130. https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C44YLT-lOAiTRKibYlV5Vjs7ioT0BO4yQ4m_mOgeS2ml3UFe8uhoTKsp8ntIwiMpi7e7k1kM8trL9wBNs0p5VD4u0&uniplatform=NZKPT&src=copy.
9. L. Liu, J. Liang, L. Ma, H.L. Zhang, Z. Lin, and S. Liang. (2022) “ Gas Pipeline Flow Prediction Model Based on LSTM with Grid Search Parameter Optimization,” Processes, 11: 63-63. 10.3390/PR11010063.
10. W.J. Wu, and J.X. Zhang. (2021) “Feature selection algorithm of random forest based on fusion of classification information and its application,” Computer Engineering and Applications, 57: 147-156. https://kns.cnki.net/kcms2/article/abstract?v=Mio27DFCfpBlVw-heCD5ebhI-j3UfIZHMKZBI30jEAy76X7s_jNK3etnRKlncfHPSSv0T5xq0Sr1Xhqp2y27iol_otFiGfte-DedVng2iHoIGsgsNzvyUtKbXLvFCv-Pg_dbxAttXmVE=&uniplatform=NZKPT&language=CHS.